The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS — Spring Semester 2016–2017

MAS474 Extended linear models — 2 hours

Restricted Open Book Examination.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator which conforms to University regulations.
Answer all questions. Total marks 60.

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 The Junior School Project collected longitudinal data on the performance of students from 49 primary schools in inner London. Over a period of 6 years, for each pupil they recorded

`school` - the school they attend (a factor with levels 1 to 49)

`class` - their class within the school (a factor with levels 1 to 4)

`gender` - the pupils gender (a factor with levels boy and girl)

`english` - their score on the English test that year (a number between 1 and 40)

`year` - the school year in which the test is taken (a number between 1 and 6)

`id` - a identification number unique to each student.

Interest lies in exploring the differences in English test performance between boys and girls. The data are stored in R as the data frame `jsp`.

(i) Initially, only the test results for year 2 are used. A model is fitted in R using the command

```
> fit1 <- lmer(english ~ gender-1+(1|school/class), data=jsp, subset = year==2)
```

If Y_{ijk} denotes the test score of student k in class j in school i, write down the equation of the model that has been used, defining your notation carefully.

(3 marks)

(ii) Explain, given the aim of the study, why the designation of explanatory variables as fixed and random is reasonable, and comment upon the structure of the random effects.

(3 marks)

1 (continued)

(iii) Use the R output below to give estimates for all of the model parameters.

```
> summary(fit1)
Linear mixed model fit by REML ['lmerMod']
Formula: english ~ gender - 1 + (1 | school/class)
   Data: jsp
 Subset: jsp$year == 2

REML criterion at convergence: 8462.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.41335 -0.75200 -0.09772  0.68060  2.90022

Random effects:
 Groups       Name        Variance Std.Dev.
 class:school (Intercept)  34.15    5.843
 school       (Intercept)  53.93    7.344
 Residual                 382.99   19.570
Number of obs: 953, groups:  class:school, 90; school, 48

Fixed effects:
           Estimate Std. Error t value
genderboy    39.751      1.581   25.14
gendergirl   44.588      1.564   28.51

Correlation of Fixed Effects:
           gndrby
gendergirl 0.652
```

(2 marks)

(iv) What is the estimated correlation between the test results for two students in the same school but different classes?

(4 marks)

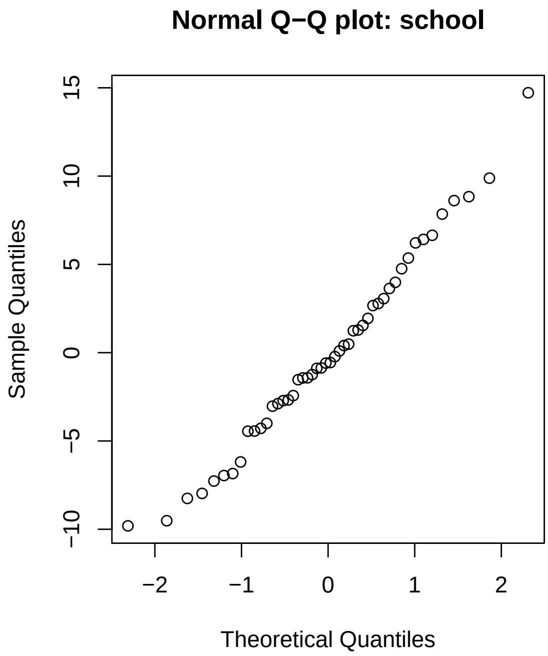
(v) The R commands

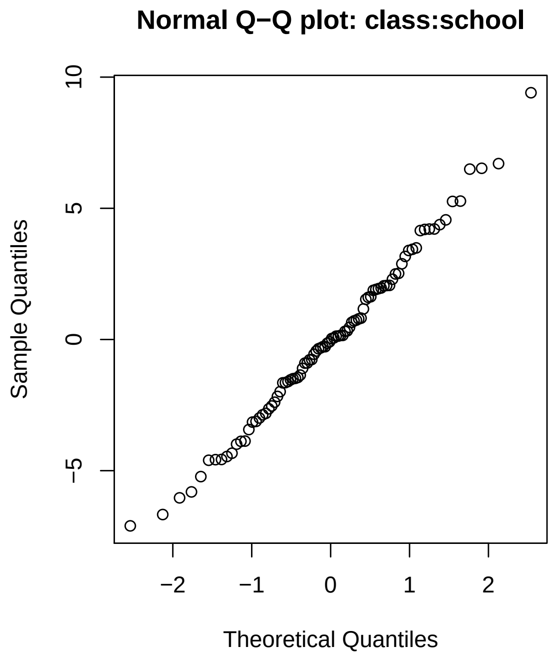
```
> par(mfrow=c(1,2))
> qqnorm(unlist(ranef(fit1)$'school'))
> qqnorm(unlist(ranef(fit1)$'class:school'))
```

were used to produce the following plots.

 Question 1 continued on next page

1 (continued)





Explain the reason for producing these plots.What do you conclude from these plots?

(2 marks)

(vi) The R session is continued below.

```
> jsp3=dplyr::filter(jsp, year==2)
> attach(jsp3)
> fit1 <- lmer(english ~ gender-1+(1|school/class), REML=F)
> fit2 <- lmer(english ~ 1+(1|school/class),  REML=F)
>
> Lambda <- replicate(10^4,{
+   english.new<-unlist(simulate(fit2))
+   fit1.new <- lmer(english.new ~ gender-1+(1|school/class), REML=F)
+   fit2.new  <- lmer(english.new ~ 1 +(1|school/class),  REML=F)
+   -2*(logLik(fit2.new) - logLik(fit1.new))
+ })
>
> Lambda.obs <- -2*(logLik(fit2) - logLik(fit1))
> sum(Lambda>Lambda.obs)
[1] 5
```

Give the name of the procedure that has been used, state what is being tested, and interpret the output.

(4 marks)

1 (continued)

(vii) The full dataset contains the test results for the pupils over their 6 years at school (longitudinal data). Write down the R command you would use to fit an appropriate mixed-effects model to investigate the trend in test performance for pupils over the 6 years they spend at primary school.

(2 marks)

2 The sequence of random variables $X_1, \ldots, X_n$ are iid with an exponential distribution with mean $\frac{1}{\lambda}$, i.e., they have pdf

$$f(x;\lambda) = \begin{cases} \lambda \exp(-\lambda x) & \text{for } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The X_i are not observed directly. Instead, we observe

$$Y_i = \mathbb{I}_{X_i \leq z_i} = \begin{cases} 1 \text{ if } X_i \leq z_i \\ 0 \text{ otherwise} \end{cases}$$

for some known sequence of thresholds $z_1, \ldots, z_n$. In other words, X_i is left-censored at z_i if $Y_i = 1$ and right-censored at z_i if $Y_i = 0$. The data can thus be summarized as $(z_1, Y_1), \ldots, (z_n, Y_n)$.

(i) Derive the likelihood function for λ given $Y_1, \ldots, Y_n$ and explain why this might necessitate the use of the EM algorithm if we want to fit the model.

(4 marks)

(ii) Derive the value for

$$R_i := \mathbb{E}(X_i | X_i > z_i).$$

Hint: recall the memoryless property of the exponential distribution. ***(3 marks)***

(iii) Show that

$$L_i := \mathbb{E}(X_i | X_i \leq z_i) = \frac{1 - (1 + \lambda z_i)e^{-\lambda z_i}}{\lambda(1 - e^{-\lambda z_i})}.$$

Hint: you may find your answer from part (i) and the law of total expectation useful

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)).$$

(6 marks)

(iv) Hence derive an EM-algorithm for finding the maximum likelihood estimator of λ using $(X_1, y_1, z_1), \ldots, (X_n, y_n, z_n)$ as the completed dataset. You may leave your answer in terms of L_i and R_i.

(7 marks)

3 (i) Consider a bivariate sample on (Y_1, Y_2) from a model that has parameters μ and ψ. For each of the following missing-data mechanisms, state whether the data are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR).

(a) $\mathbb{P}(y_2 \text{ missing } | y_1, y_2, \mu, \sigma, \psi) = \dfrac{e^{\mu+\psi y_1}}{1 + e^{\mu+\psi y_1}}$

(b) $\mathbb{P}(y_2 \text{ missing } | y_1, y_2, \mu, \sigma, \psi) = \dfrac{e^{\mu+\psi}}{1 + e^{\mu+\psi}}$

(c) $\mathbb{P}(y_2 \text{ missing } | y_1, y_2, \mu, \sigma, \psi) = \dfrac{e^{\mu+\psi y_1 y_2}}{1 + e^{\mu+\psi y_1 y_2}}$

Which of these mechanisms are ignorable for likelihood-based inference?

(4 marks)

(ii) Consider the following dataset, where `NA` is used to denote a missing observation.

x	y
-6.3	-9.5
1.8	6.4
NA	-13.7
16.0	NA

Interest lies in the expected value of y. Calculate the mean of y using

(a) Complete-case analysis

(b) Available-case analysis

(c) Mean imputation

(d) (Mean) Regression imputation

You may find the following R output useful.

```
> coef(lm(y~x))
(Intercept)           x
   2.866667    1.962963
> coef(lm(x~y))
(Intercept)           y
  -1.460377    0.509434
```

(5 marks)

 Question 3 continued on next page

3 (continued)

(iii) Consider a dataset from a large postal survey on the psychology of debt. It contains responses from $n = 464$ individuals. The covariates are

- `AttDebt` - score on a scale of attitudes to debt (high values=favourable to debt)
- `CCard` - how often did s/he use credit cards (1=never... 3=regularly)
- `BuildSoc` - does the respondent have a building society account?
- `LOC` - the respondents' *locus of control*, which is the degree to which the respondent believes that they have control over the outcome of events in their life (high values = belief that one's life can be controlled).

The output below shows the structure of the dataset.

```
> str(Debt)
'data.frame': 464 obs. of  4 variables:
 $ AttDebt : num  2.71 3.88 3.06 4.29 3.82 ...
 $ CCard   : Factor w/ 3 levels "1","2","3": 2 3 2 2 3 2 2 1 NA 2 ...
 $ BuildSoc: Factor w/ 2 levels "0","1": NA NA NA 1 1 1 1 1 NA ...
 $ LOC     : num  2.83 4.83 3.83 4.83 3.17 ...
> head(Debt)
  AttDebt CCard BuildSoc  LOC
1    2.71     2     <NA> 2.83
2    3.88     3     <NA> 4.83
3    3.06     2     <NA> 3.83
4    4.29     2        0 4.83
5    3.82     3        0 3.17
6    3.06     2        0 3.83
```

(a) The following R command is used.

```
> Debt.mice <- mice(Debt, m=5, method = c('norm', 'polyreg', 'logreg', 'norm'))
```

Describe in detail the statistical procedure that is used to fill in the missing values.

(6 marks)

3 (continued)

(b) Interest lies in the coefficient of LOC (β_3) in the linear regression model

$$\texttt{AttDebt} = \beta_0 + \beta_1 \texttt{CCard} + \beta_2 \texttt{BuildSoc} + \beta_3 \texttt{LOC}$$

Use the R output below to calculate an expected value of β_3 and its standard error.

```
> fit.mice <- with(Debt.mice, lm(AttDebt ~ CCard+BuildSoc+LOC))
> (coefs = sapply(fit.mice$analyses, coef))
             [,1]  [,2]  [,3]  [,4]  [,5]
(Intercept)  3.61  3.74  3.73  3.68  3.75
CCard2       0.11  0.15  0.17  0.19  0.12
CCard3       0.45  0.45  0.41  0.39  0.45
BuildSoc2   -0.21 -0.19 -0.16 -0.13 -0.16
LOC         -1.12 -2.11 -0.53 -1.32 -1.01
>
> apply(coefs, 1, var)
(Intercept)      CCard2      CCard3   BuildSoc2         LOC
    0.00337     0.00112     0.00080     0.00095     0.33307
>
> sapply(fit.mice$analyses, function(x) vcov(x)[5,5])
[1]  0.67 0.78 0.46 0.90 0.46
```

(5 marks)

End of Question Paper